Unsupervised Sense Classification For Word Sketches

Ondřej Herman

Lexical Computing
ondrej.herman@sketchengine.eu

Introduction

- Enriching syntax-based Word Sketches with word sense information.
- Sketch Engine Feature in development.

- How do Word Sketches work?
- Word Sense Induction.
- How to combine these?

WORD SKETCH

English Web 2020 (enTenTen20)

Get more space

bank as noun 4,676,470×

modifiers of "bank"		
central	190,112	10.3
the central bank		
• especially: business		
investment	40,826	8.7
investment banks		
• especially: business		
river	31,710	8.6
the river bank		
• especially: recreation		
commercial	43,750	8.2
commercial banks		
• especially: business		
food	42,581	7.9
food banks		
west	17,901	7.9
on the west bank of the		
north	14,234	7.5

nouns modified by "bank"		
account	224,638	10.7
bank account		
• especially: business		
robbery	14,986	8.7
a bank robbery		
loan	25,946	8.4
bank loans		
• especially: business		
robber	11,371	8.4
bank robber		
holiday	19,710	8.3
bank holidays		
• especially: business		
transfer	18,106	8.1
bank transfer		
• especially: business		
deposit	13,671	8.0

verbs with "bank" as object		
rob	15,354	9.2
rob a bank		
break	28,294	8.0
without breaking the bank		
overflow	3,358	7.2
overflowed its banks		
burst	3,069	7.0
burst its banks		
nationalize	2,843	7.0
nationalized banks		
• especially: business		
nationalise	2,400	6.7
nationalised banks		
line	3,001	6.6
line the banks		
• especially: recreation		
own	7,662	6.5

Word Sketch

- Extracts frequent patterns from corpus data.
- Expert-designed rules.

```
=modifier
2:"(JJ|NN).?" [tag="JJ.?"|tag="RB.?"|word=","]{0,3} "NN.?.?"{0,2}   1:"NN.?.?"
[tag!="NN.?.?"]

2:"RB" 1:[tag="JJ.?"|tag="V..?"]
```

- Competition among foreign commercial banks has resulted in significant
- Competition among foreign commercial banks has resulted in significant
- Situated on the western bank of the Volga River
- central bank utilize a capital buffer for large banks to counteract potential losses

WORD SKETCH

English Web 2020 (enTenTen20)

Get more space

bank as noun 4,676,470×

modifiers of "bank"		
central	190,112	10.3
the central bank		
• especially: business		
investment	40,826	8.7
investment banks		
• especially: business		
river	31,710	8.6
the river bank		
• especially: recreation		
commercial	43,750	8.2
commercial banks		
• especially: business		
food	42,581	7.9
food banks		
west	17,901	7.9
on the west bank of the		
north	14,234	7.5

nouns modified by "bank"		
account	224,638	10.7
bank account		
• especially: business		
robbery	14,986	8.7
a bank robbery		
loan	25,946	8.4
bank loans		
• especially: business		
robber	11,371	8.4
bank robber		
holiday	19,710	8.3
bank holidays		
• especially: business		
transfer	18,106	8.1
bank transfer		
• especially: business		
deposit	13,671	8.0

verbs with "bank" as object		
rob	15,354	9.2
rob a bank		
break	28,294	8.0
without breaking the bank		
overflow	3,358	7.2
overflowed its banks		
burst	3,069	7.0
burst its banks		
nationalize	2,843	7.0
nationalized banks		
• especially: business		
nationalise	2,400	6.7
nationalised banks		
line	3,001	6.6
line the banks		
• especially: recreation		
own	7,662	6.5

Word Sketch

- Easy to interpret.
- Easy to understand.
- The result is supported by corpus evidence.
  - You can see all the underlying instances.

- Purely syntax based.

Word Senses

- What is a Word Sense?
- Distributional Hypothesis.
    - Words appearing in similar contexts tend to have similar meanings.

Word Sense Induction

- We have explored many methods over the years.
  - Dictionary drafting.
- Simple solutions do not seem to work well or have many parameters.
- Getting senses which correspond to human intuition is hard.
- AI complete?

Adaptive SkipGram

- Word Sense Induction algorithm.
  - *Bartunov, S., Kondrashkin, D., Osokin, A., & Vetrov, D. (2016, May). Breaking sticks and ambiguities with adaptive skip-gram. In artificial intelligence and statistics (pp. 130-138). PMLR.*
- Embedding based.
- Speed of fastText (per sense).
- For every word, multiple vectors are learned, one per sense.
  - Subject to parameter α, granularity of the resulting senses.
- The model is precomputed on the whole corpus text.
  - The model describes the senses present in the corpus.
  - The model is able to desambiguate words based on contexts.

Adaptive SkipGram

- The original implementation is written in Julia.
- We rewrote the algorithm in the Rust language.
  - Performance, maintainability.

- The model is difficult to interpret.

Mouse

Desambiguating the Word Sketch

- The sense is determined for every WS item.
  - Triple consisting of (headword, grammatical relation, collocate).
- For every instance, the sense is desambiguated by the WSI model:

s, inundating many outlying localities and villages situated on the **river**	**bank**	, affecting about one lakh people, flood waters of the Gomti entered the Vipul
volunteers removed over 46 tons of trash from over 249 miles of **river**	**banks**	and waterways.</s><s>To give you a sense, 46 tons is equal to 322,000 apple
.</s><s>Used syringes, infectious bandages were all around the **river**	**banks**	.</s><s>After experiencing these haunting sites, Toxics Link decided to deal w
es, and those that broke and fled were shot as they ran along the **river**	**bank**	or in the open towards Warneton.</s><s>Among the decorations bestowed on
><s>A party, estimated at about fifty, collected under cover of the **river**	**bank**	, and made towards our right flank.</s><s>Nicol at the moment was near the
er (You will most likely buy fish from the villager for a lunch on the **river**	**bank**	!).</s><s>The boat trip continues all the way to Vietnamese border and it take
s><s>After passing over this bluff the trail runs along a flat on the **river**	**bank**	till another precipice forces it into the stream when it is carried along the base

- This yields a probability distribution over the word senses.
- The average probability per sense is calculated.

Evaluation

- Manually crafted test set for English.
- Known-polysemous words.

band bank bark base bat beam board bow change chip club
crane deck file iron jumper party pole spring tank tear

- For each word, 150 top elements by association score were annotated for word sense.

Evaluation

- Recall = proportion of senses in the test set found by the WSI algorithm.
  - How many of the annotated senses were found automatically?
- Precision = proportion of senses found by the WSI algorithm in the test set.
  - How many of the automatically found senses were present in the test set?

- It is possible to trade recall for precision during the training step.

Evaluation

- Evaluated against a 500 M token corpus sample.
- Up to 70 % F1 score.
- 77 % Recall, 62 % Precision (2 train epochs, context size 10, dim 128)

- Almost no change in performance w.r.t. dimensionality (64, 128, 256).
- Context sizes over 10 do not help much.

Examples

Crown

Blob

Crane

Test

Bank

Mouse

https://projects.sketchengine.eu/eca3eee0

Future Work

- Granularity of the induced senses.
  - Different parametrization yields different
- How to deal with overlapping senses?
- Improve sense desambiguators.

Conclusion

- We are enriching syntax-based word sketches with word sense information.
- 70 % F1 against hand-crafted test set.

https://projects.sketchengine.eu/eca3eee0

Thank you!

Questions?

https://projects.sketchengine.eu/eca3eee0

ondrej.herman@sketchengine.eu